ON THE HOMOGENEITY OF SEQUENTIAL QUOTA-SAMPLES OF CLUSTERS: A PARAMETER-FREE DISTRIBUTION

BU-601-M January 1977

C. L. Wood and D. S. Robson

Abstract

Sample size is sometimes fixed in terms of total bulk or some other total size quota on the sampled units rather than on the total number of units. When the sampling units are households, for example, the sampling design might specify the total number of people rather than the total number of households to be included in the sample. Sampling of households then terminates when the accumulated number of people in the sample reaches this predetermined limit, with possibly a slight excess over the limit due to the size of the terminal household.

In such cases where a limit is placed upon the sample total of a positive discrete random variable, the sample frequency distribution is a sufficient statistic. When k independent samples are available then the frequency distribution in the pooled sample is sufficient with respect to the hypothesis that all samples were drawn from the same population. The joint probability distribution of the k sample frequency distributions, conditioned upon the pooled frequencies and the k predetermined boundary values, is then parameter-free and provides a basis for constructing statistical tests of homogeneity.

ON THE HOMOGENEITY OF SEQUENTIAL QUOTA-SAMPLES OF CLUSTERS: A PARAMETER-FREE DISTRIBUTION

BU-601-M C. L. Wood and D. S. Robson January 1977

1. Introduction

The existence of an exact small-sample test of homogeneity of r multinomial samples derives from the sufficiency of the class frequencies in the combined sample. Conditional upon the total class frequencies, as well as upon the (predetermined) sample sizes, the H_0 distribution of individual sample frequencies is parameter-free – i.e., is functionally dependent only upon available or observable parameters. We demonstrate here that a similar argument applies to r sequential samples from the same discrete p.m.f. on the non-negative integers, when the stopping rule is defined by a bound t_i on the total of the counts in the i^{th} sample. Such rules might apply, for example, when the counted items must be stored for processing and t_i is the storage or processing capacity available for the i^{th} sample, or when items are sampled by clusters and t_i is the quota on the number of items in the i^{th} sample.

2. A Parameter-Free Distribution

A single sequential sample from the probability mass function (p.m.f.)

$$f_Y(j) = P(Y=j) , \qquad j=0,1,2,\cdots$$

terminates at the first value of n for which the sum $S = Y_1+\cdots+Y_n$ equals or exceeds a predetermined bound $t > 0$. Occurrences of $Y_i = 0$ in this sequence represent only a nuisance complication in the present analysis so we assume $f_Y(0) = 0$ *pro tempore*.

The likelihood L_t of any particular sequence $(y_1,\cdots,y_n)$ satisfying this stopping rule is

$$L_t(y_1,\cdots,y_n) = \prod_j [f_Y(j)]^{x_j}$$

and the vector of "class frequencies" x_j,

$$x_j = \#\{\nu | 1 \le \nu \le n,\ y_\nu = j\}\ ,$$

with

$$\sum_j x_j = n$$

and

$$\sum_j jx_j = s \ge t$$

is seen to be sufficient with respect to the p.m.f. $f_Y(j)$ with $f_Y(0) = 0$.

The p.m.f. of this sufficient statistic $\underset{\sim}{X}$ is obtained by summing the likelihood L_t over all sequences $(y_1,\cdots,y_n)$ producing the class frequencies $\underset{\sim}{x}$ and satisfying the stopping rule defined by the specified t. Letting $C_t(\underset{\sim}{x})$ denote the number of such sequences, then

$$P_t(\underset{\sim}{X} = \underset{\sim}{x}) = C_t(\underset{\sim}{x})L_t \qquad (1)$$

and this combinatoric coefficient is readily found to be

$$C_t(\underset{\sim}{x}) = \frac{n!}{\prod_j x_j!}\left[\frac{1}{n}\sum_{j=s-t+1} x_j\right].$$

If the matrix $\bar{\underset{\sim}{X}}' = [\underset{\sim}{x}_1', \cdots, \underset{\sim}{x}_r']$ denotes the outcome of r sequential samples selected independently from this same population but with stopping rules given

by $\underset{\sim}{t} = (t_1, \cdots, t_r)$, respectively, then the joint p.m.f. of $\bar{\underline{X}}$, $P_{\underset{\sim}{t}}(\bar{\underline{X}})$, will be the product of (1) over the r samples. The Neyman factorization theorem reveals that the vector of class frequencies for the combined sample is again sufficient with respect to $f_Y(j)$, $f_Y(0) = 0$; thus,

$$P_{\underset{\sim}{t}}(\bar{\underline{X}}) = C_{\underset{\sim}{t}}(\bar{\underline{X}}) \prod_j [f_Y(j)]^{x_{\cdot j}}$$

where $\underset{\sim}{x}_\cdot$ is the vector of class frequencies in the combined sample,

$$x_{\cdot j} = \sum_{i=1}^{r} x_{ij} ,$$

and

$$C_{\underset{\sim}{t}}(\bar{\underline{X}}) = \prod_{i=1}^{r} C_{t_i}(\underset{\sim}{x}_i) .$$

The parameter-free p.m.f. of $\bar{\underline{X}}$ conditional on the sufficient statistic $\underset{\sim}{x}_\cdot$ is therefore given by

$$P_{\underset{\sim}{t}}(\bar{\underline{X}} | \underset{\sim}{x}_\cdot) = \frac{C_{\underset{\sim}{t}}(\bar{\underline{X}})}{\sum_{\bar{\underline{X}} | \underset{\sim}{x}_\cdot} C_{\underset{\sim}{t}}(\bar{\underline{X}})} , \qquad (2)$$

where the sum in the denominator extends over all $\bar{\underline{X}}$-matrices having column sums given by $\underset{\sim}{x}_\cdot$ and having rows satisfying the r stopping rules given by $\underset{\sim}{t}$.

3. Discussion

The above results constitute a basis for formulating a research and development problem in statistical methodology. The parameter-free conditional p.m.f. (2) is a sequential sampling analogue to the multihypergeometric distribution

which would have arisen in the more conventional case of sampling with predetermined sample sizes $n_1,\cdots,n_r$. Methodology employed in the latter, familiar situation will suggest analogous methods to be developed for the sequential case. Construction of a test statistic analogous to the classical, contingency chi-square statistic, for example, would require calculation of first and second order moments of the exact or of an asymptotic version of (2). Such results might also be used in constructing an analogue to the conventional randomization test statistic for testing homogeneity among the means of r samples of (large) fixed sizes.

Justification and ultimate use of such test statistics would rest primarily upon asymptotic theory supported by numerical comparisons between exact and asymptotic distributions. Algorithms are therefore required for calculating (2) as well as moments and distributions derived from (2). The joint and marginal distributions of the $\underset{\sim}{n} = (n_1,\cdots,n_r)$ and $\underset{\sim}{s} = (s_1,\cdots,s_r)$ vectors, for example, might well be required in the development of test procedures.

One modification which might be required for analytic or computational convenience is the censoring of the terminal observation y_n in any sample producing an excess over the boundary. This censoring may be accomplished notationally by letting

$$(n_i^*,s_i^*) = \begin{cases} (n_i,s_i) & \text{if } s_i = t_i \\ (n_i-1,\ s_i-y_{i,n_i}) & \text{if } s_i > t_i \end{cases} ,$$

$$x_{ij}^* = \#\{\nu \mid 1 \le \nu \le n_i^* ,\ y_{i\nu} = j\} ,$$

and the conditional p.m.f. of the censored data is then given by (2) with asterisks on all x_{ij}. The potential convenience deriving from this type of censoring is indicated by

$$P_{\underset{\sim}{t}}(\bar{\underset{\sim}{X}}^{*}|\underset{\sim}{x}_{\cdot}^{*}) = P_{\underset{\sim}{s}^{*}}(\bar{\underset{\sim}{X}}^{*}|\underset{\sim}{x}_{\cdot}^{*}) \quad ,$$

where $C_{s^{*}}(\bar{\underset{\sim}{X}}^{*})$ then has the simplified form

$$C_{s^{*}}(\bar{\underset{\sim}{X}}^{*}) = \prod_{i=1}^{r} \left(\frac{n_{i}^{*}!}{\prod_{j} x_{ij}^{*}!} \right) \quad ,$$

with

$$\sum_{j} x_{ij}^{*} = n_{i}^{*}$$

and

$$\sum_{j} j x_{ij}^{*} = s_{i}^{*} \quad .$$

There might, indeed, be applied circumstances where real censoring is either convenient or unavoidable, and so require the development of methodology for censored data.

An extension which might be required for some applied settings is inclusion of the possibility $y = 0$; i.e., $f_{Y}(0) > 0$. In a number-quota fishing operation, for example, where y_{ν} is the number of fish captured in the ν^{th} net haul, the event $y_{\nu} = 0$ might occur with positive probability. In other types of quota-sampling, including bulk-sampling, the non-negative discrete variable y might not be integer-valued, as in a volume-quota fishing operation on a (large) fish population composed of discrete size classes corresponding to non-overlapping age groups. In such cases where the p.m.f. of y has support on the positive values $0 < \lambda_{1} < \lambda_{2} < \cdots$, the result (2) still holds with $x_{ij} = \#\{\nu | 1 \leq \nu \leq n_{i}, y_{i\nu} = \lambda_{j}\}$ and again with $s_{i} = y_{il} + \cdots + y_{in_{i}}$. Inclusion of zeroes in this more general model would be required for some applications.